Cureus Journal of Computer Science

Open Access Technical Report

Artificial Intelligence and Privacy Concerns: Balancing Innovation With Security

Vaibhav Viswanath [1] , Thenmozhi M [2] , Sahana Naganandh [1]

1. School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, IND 2. School of Social Sciences and Languages, Vellore Institute of Technology, Vellore, IND

Corresponding author: Thenmozhi M, thenmozhi.nisha@vit.ac.in

Received 03/10/2025
Review began 04/17/2025
Review ended 04/21/2025
Published 04/24/2025

DOI:
https://doi.org/10.7759/s44389-025-03689-z

Abstract

This report investigates the issues of privacy concerning multilingual artificial intelligence (AI) datasets and investigates the functionality of PII43K as an alternative for privacy preservation. AI models trained on multilingual datasets, often inclusive of English, French, German, and Italian languages, risk unauthorized access, theft, and regulatory noncompliance with sensitive levels of personally identifiable information (PII). Using a structured approach encompassing data preprocessing, named entity recognition, token count analysis, and risk comparison methods, the research attempts to evaluate the extent of PII exposure and the efficacy of PII43K to mitigate such risks. The results show that traditional multilingual datasets present major privacy vulnerabilities due to direct injection of such sensitive information, while PII43K safely anonymizes its data using masking, redaction, and structured tokenization. Furthermore, compliance analysis confirms that PII43K meets the requirements of both General Data Protection Regulation and California Consumer Privacy Act and is, therefore, a useful option for privacy-centric AI applications. The need for privacy-aware AI models that preserve the utility of data and simultaneously protect it and ensure compliance with regulations is underscored. Future research should broaden the scope to include languages outside Europe, explore the privacy risks of AI in domain-relevant datasets in healthcare and finance, and develop a feasibility study for the economics of large-scale AI for privacy preservation.

Categories: Ethical AI and Responsible Technology, Security and Privacy, Data Ethics
Keywords: data privacy, personally identifiable information (pii), multilingual ai datasets, privacy-preserving ai, ai model memorization, secure ai training, regulatory compliance

Introduction

Background of the study

Enhanced trust in personal data use in artificial intelligence (AI)-assisted applications depends on balancing privacy and identity protection across different domains. AI presents various dilemmas regarding privacy, including bias and discrimination, raising critical questions about user consent, accessibility, ownership, and misuse deterrence [1]. AI functions as both a recognition and prediction tool, introducing ethical concerns about how personal information is stored, accessed, and processed [2]. The increasing reliance on AI amplifies risks such as unauthorized access, data security breaches, discrimination, and algorithmic bias, making privacy and security central to AI governance discussions [3].

To address these concerns, AI development must be framed within legally substantive frameworks that enforce ethical principles [4]. The General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) provide governance models for responsible AI implementation. These regulations emphasize transparency, accountability, and user control, ensuring that AI applications respect data privacy [5]. Despite the presence of regulatory mechanisms, gaps still exist, making AI training and decision-making processes vulnerable to ethical violations.

AI introduces both solutions and risks in various domains, particularly in cybersecurity. AI-powered security systems enhance threat detection, fraud prevention, and risk assessment, safeguarding digital infrastructure [6]. However, AI also facilitates cyberattacks, including adversarial attacks and deepfake manipulation, challenging encryption protocols and cybersecurity defenses [7]. Such vulnerabilities necessitate the continuous development of robust AI security measures [8].

The rapid transformation of industries due to AI has driven research toward privacy-preserving AI technologies, which aim to balance innovation with data protection [9]. Techniques such as differential privacy, federated learning, and secure multi-party computation enable AI systems to operate without compromising sensitive data [10]. Additionally, explainable AI has emerged as a key solution for ensuring transparency, accountability, and fairness in AI decision-making processes [11].

Given the challenges and opportunities AI presents, companies must prioritize privacy, ethical

governance, and regulatory compliance. This research investigates AI's privacy implications, identifies regulatory bottlenecks, and explores future technological advancements in responsible AI governance. Additionally, it highlights industry best practices for privacy-preserving AI, contributing to the development of an ethical AI framework emphasizing fairness, robustness, and transparency. Ultimately, this study provides insights into AI regulations that support a secure and privacy-conscious technological future.

Relevance of study of AI has rapidly transformed industries through automation, decision-making, and operational efficiencies. However, this growth has raised concerns about data privacy, security, and ethical obligations [12]. AI systems rely on vast amounts of personal and sensitive data, intensifying concerns over data collection, storage, and usage [13]. Balancing AI-driven innovation with privacy protection is a critical challenge to ensure that technological advancements do not infringe on individual rights [14].

AI's ability to approximate, predict, and manipulate user behavior - sometimes without explicit consent - raises ethical concerns [15].

Issues such as mass surveillance, opaque algorithmic processes, and data breaches highlight the urgency of establishing regulatory frameworks and ethical AI deployment practices.

While regulations like GDPR and CCPA aim to mitigate privacy risks, they remain insufficient in counteracting the full extent of AI's data-driven capabilities. This study seeks to bridge the gap between innovation and privacy protection by analyzing threat vectors, legal frameworks, and technical countermeasures such as differential privacy and federated learning [10].

Objectives of the study

-To analyze the impact of AI on data privacy by identifying key risks related to data collection, surveillance, and algorithm-based decision-making.

-To evaluate privacy threats in multilingual AI datasets, particularly the exposure of personally identifiable information (PII) such as names, emails, financial data, and tracking details.

-To investigate the ethical concerns of AI-based decision-making, with a focus on bias, consent, and transparency in data processing.

-To assess the effectiveness of privacy-preserving techniques, such as PII masking, redaction, and structured anonymization, in mitigating AI privacy risks.

-To analyze industry best practices and regulatory compliance measures, particularly adherence to GDPR and CCPA standards in AI privacy governance.

-To provide recommendations for future AI privacy frameworks, emphasizing the balance between technological progress and data security.

Rationale of the study

The justification for this study lies in the growing role of AI in processing vast amounts of user data, raising concerns about privacy, security, and ethical governance [16]. AI-driven systems influence decision-making processes in sectors such as finance, healthcare, and law enforcement, necessitating strict data protection mechanisms [17].

The ability of AI to collect, analyze, and infer sensitive information raises ethical concerns about data usage, compliance with regulations, and cybersecurity risks [6].

This research systematically synthesizes existing literature on AI privacy risks, security challenges, and regulatory frameworks. It contributes to the field by exploring AI governance frameworks that balance innovation, privacy, and regulatory compliance; assessing privacy-preserving AI techniques, such as differential privacy and federated learning [10]; and identifying key privacy risks in multilingual AI datasets [18]. By addressing these issues, this study aims to provide valuable insights into privacy-conscious AI development, ensuring that AI innovation progresses without compromising individual rights or security.

Technical Report

Natural language processing (NLP) and machine learning models are usually trained using multilingual datasets that consist of text in English, French, German, and Italian. These datasets contain text taken from social media, legal documents, and customer interactions to improve the AI's language-processing

capabilities. Unfortunately, these datasets come with privacy concerns since they also contain PII, such as names, email addresses, IP addresses, and financial information.

AI models can easily learn and reproduce sensitive data in the absence of proper anonymization, including identity theft and infringements on privacy regulations, such as GDPR and CCPA. To resolve these issues, the PII43K dataset was created to ensure the protection of privacy by way of PII masking and redaction. This paves the way for secure AI training, free of any risk of exposing actual user data, through the use of PII43K [19]. The use of PII masks and redactions thus permits harmless AI training, posing a lower risk to organizations that primarily prioritize data security and compliance with established regulations.

Methodology

Data Sources

This study employs several datasets in analyzing privacy risks and investigates whether PII43K is found to mitigate these risks in action. The research will mostly benefit from the primary datasets, listed as follows:

Multilanguage PII Datasets: This dataset includes English, French, German, and Italian data containing PII, covering various sensitive topics such as names, dates, locations, email addresses, financial data, and user tracking information, all captured through open internet channels.

To reduce the risk of disclosure of PII, the primary datasets are privacy sensitive and efficient.

The PII43K Dataset: Whilst ensuring that data is meaningless for AI applications, it has been produced in a way that can still maintain the utility of privacy. Such structured anonymization techniques are composed of entity masking and redaction.

Raw evaluation will be given to its multilingual datasets for framing the privacy risks, also considering how well PII43K works in the context of mitigating those risks. It is hoped that data pertaining to privacy risks be captured in benchmark and the benefits of structured anonymization of these virtual datasets be analyzed.

Data Collection

Data collection will be done in a systematic manner to ensure that accuracy and parity in analysis are successfully reflected. The following steps would be implemented:

Dataset Acquisition: Datasets will be found from open-source repositories and research archives in English, French, Italian, and German. Then, PII43k dataset and private AI repositories would be obtained [19].

Pre-Processing and Data Cleaning: The dataset will be inspected for inconsistencies, missing values, or double entries. The uniformity should be ensured addressing format errors, erring of delimiters, encoding mismatches, and improperly structured data. Any other unwanted noise in these datasets, which could affect the accuracy of PII discovery, will be effectively eliminated.

Tokenization and Text: Processing Breaking texts into individual tokens would be the first step for interpreting the data, which comprises words, numbers, and symbols. Thus, we will analyze the composition of PII measures within the datasets. NLP would be put to work using named entity extraction and the labeling of these kinds of entities into various PII categories.

Tagging and Classifying of PII: Every dataset will undergo a process of marking out any PII using automatic machine learning models centered on named entity recognition. Stakeholders will be responsible for monitoring to enhance accuracy and reduce mistakes in labeling.

Analytical Methods

To assess the privacy threat level and PII disclosure according to certain criteria, we will employ multiple analytic approaches:

Token Count Analysis: This will measure the frequency of PII tokens in each dataset, indicating how often different personal identifiers appear.

Category-Wise PII Distribution: We will categorize PII elements (e.g., name, email, IP, financial information) to identify areas with higher risks.

Comparable Risk Assessment: The proportion of PII tokens relative to the total dataset size will be calculated. A comparison will be built upon the density of exposed PII in different languages.

Graphical and Statistical Analysis: We will use bar charts, histograms, and pie charts to visually represent trends in PII exposure.

Inference of Statistical Significance: We will draw conclusions based on statistical validation to establish the significance of our findings.

Evaluation of PII43K Dataset

To determine the effectiveness of PII43K in allaying privacy concerns, we establish a comparison between its structure, content, and language datasets [19]. The evaluation will investigate:

PII Exposure Reduction: Characterization of the proportion of masked or redacted PII entities in PII43K compared to raw multilingual datasets. The anonymization success index will show, on a percentage scale, how many direct identifiers were obliterated.

Data Utility Retained: Evaluation of how well PII43K maintains essential data characteristics without compromising privacy. Context-aware evaluation techniques will be exercised in determining the usefulness of anonymized data for AI training.

Cross-Language Performance: Consistency assessments across different languages to determine whether inquiries to this data repository effectively obviate any PIIs. They ascertain differences in privacy risks between different languages to document the adaptiveness of the data for graduation with multilingual AI models.

Regulatory Compliance Review: Assessment of PII43K against GDPR and CCPA compliance to ensure its validity with legal standards. Besides this, an ethical AI development perspective will be presented that concerns itself with the proper handling of the data.

Under this methodology, the findings will document all privacy risks for multilingual datasets, while demonstrating the benefits of PII43K as a clear alternative for privacy. This will help build privacy and necessity models for AI.

Overview of multilingual AI datasets

These datasets are named English, French, German, and Italian, and are commonly used for training and testing AI systems in activities such as NLP, text automation, and machine learning. They contain massive amounts of real-life textual data, making them an ideal training ground for AI models. However, they also pose significant privacy risks due to the vast amounts of PII they contain.

Characteristics of the Multilingual Datasets

Multilingual AI Training: These datasets contain data in English, French, German, and Italian that allow AI models to process and understand multiple languages.

Rich Data Sources: This includes social media texts, online articles, legal documents, and customer interactions among many others to make up that diverse dataset for AI learning.

High Presence of PII: These datasets contain sensitive user data, such as Names (First Name, Last Name, Full Name); Email Addresses and Phone Numbers; IP Addresses and User Agents (Tracking Data); Financial Data (e.g., IBANs, Bitcoin and Ethereum Wallets).

Privacy risks associated with multilingual datasets

From a multilingual dataset analysis (English, French, German, and Italian), it was found that the real high risks of exposure come from PII. Accordingly, the following observations were made as per graphical analysis (Figure *1*) [19].

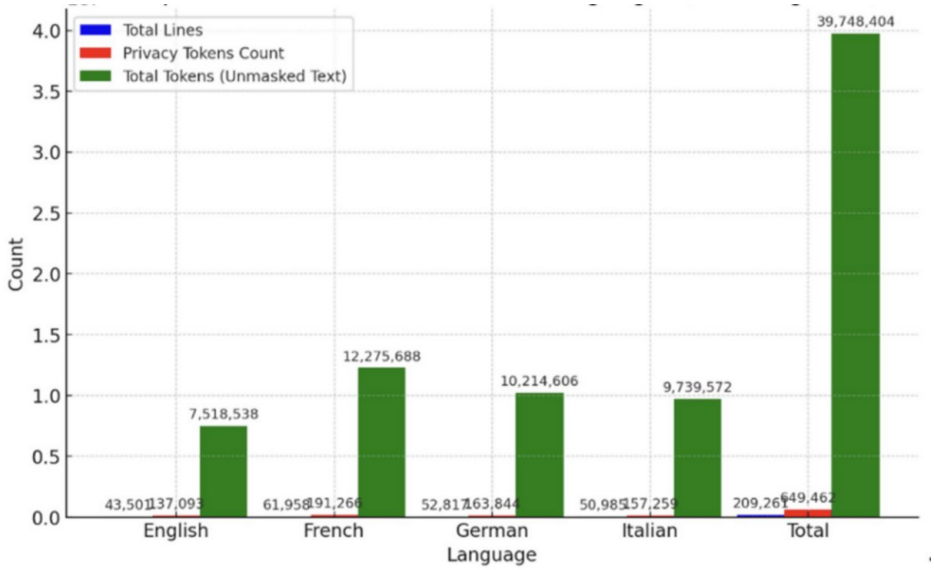

FIGURE 1: Comparison of Data Metrics Across Languages (Including Total).

Blue - Total Lines; Red - Privacy Tokens Count; Green - Total Tokens (Unmasked Text)

High Exposure of PII Categories

- For all languages, the first names stand out in exposure, and this is even more with the French data (with an incidence of 20,081), followed by the German and the Italian (Figure *2*).

- Among others that are highly exposed are last names and dates, emails, and financial data (Bitcoin addresses, IBANs, and IPs).

- The IP presence indicates the likelihood of tracking and potential risks from surveillance over all these databases.

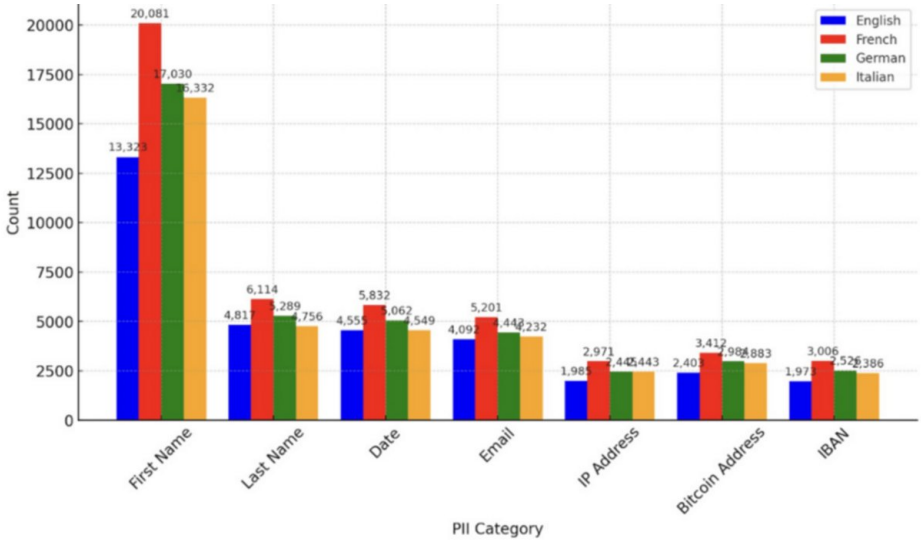

FIGURE 2: PII Exposure Across Languages.

Blue - English; Red - French; Green - German; Yellow - Italian

PII, Personally Identifiable Information

Disproportionate Representation of Sensitive Data

- The French dataset stands alone with a very high proportion of privacy tokens (Figure 3), having counted 1,91,266 tokens and giving AI models even more opportunity for personal data retention.

- While the English set carries less total lines than the French one (Figure 3), it carries too much privacy-sensitive information, thus leading to an increase in possibilities for privacy infringement.

- Overall, there are over 6,49,000 instances of privacy tokens available in the datasets, suggesting that the sensitive data are, for the most part, available and unprotected.

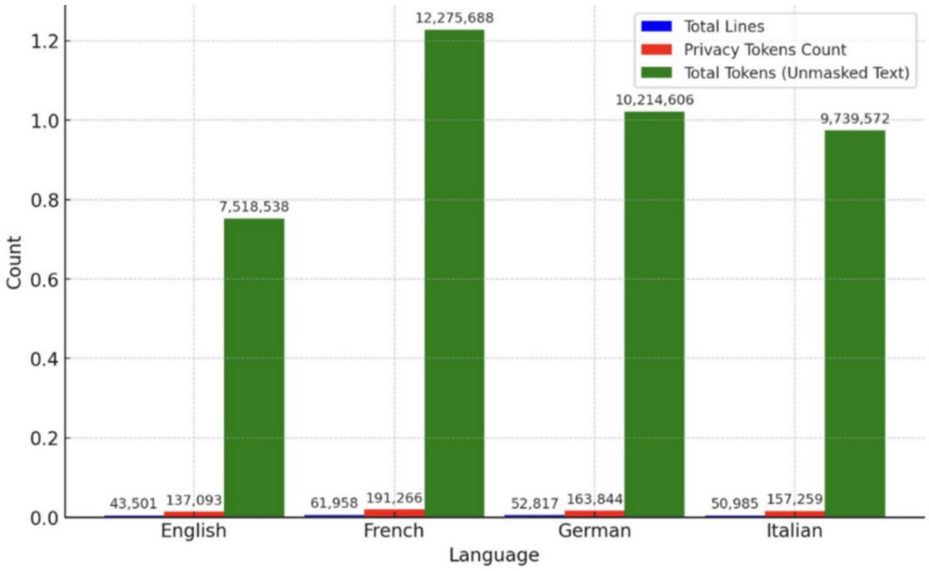

FIGURE 3: Comparison of Data Metrics Across Languages.

Blue - Total Lines; Red - Privacy Tokens Count; Green - Total Tokens (Unmasked Text)

Risk of Regulatory Non-Compliance

- Such hardly qualifies for privacy protection, meaning that should AI actually employ such datasets for its training, it would have no defence against GDPR and CCPA.

- The absence of anonymization and redaction can give the models a chance of spoiling the release of sensitive data.

Risk of AI Model Memorization and Data Leaks

- PII in unfiltered datasets can be virtually memorized by AI models.

- Use of structured obfuscation will prevent the possible duplication of sensitive information, limiting the appearance of ethical and legal issues.

Key Takeaway

The multilingual datasets under study are putting much privacy at stake because of PII exposure, non-anonymization, and non-compliance with relevant regulations abound. This highlights the importance of privacy-controlled datasets such as PII43 for avoiding these risks and increasing safety for AI.

While significant for AI training, these datasets are great threats to privacy in their raw, unrefined states.

Identified privacy risks in multilingual datasets

Direct Exposure of Personal Data

It has certain financial details, such as real name, email, and financial data, in an actual, non-anonymous form. Thus, AI models may memorize these data and repeat them again by mistake.

Tracking & Surveillance Risks

In the case of this data, obtaining source details is not only limited to the IP addresses but is extended to include the user agent information, meaning they can unlawfully spy on the users. Such AI systems can easily figure out the user's browsing preferences and the device used by the respective user to infringe the privacy of the user.

Risk of Identity Theft & Fraud

The availability of such details as financial data (e.g. IBAN, Bitcoin, Ethereum) makes it easily prone to fraudulent activities. One of the consequences of training such AI systems with such data can be the unintentional loss or circulation of the above mentioned financial details as such data applicable in response generation is used as a parameter.

Lack of Privacy Controls

These datasets are not in compliance with privacy laws, be it GDPR (EU) or CCPA (California), for example. Besides, they are not designed for any privacy-concerned AI applications because they do not use masking or redacting.

Real-World Implications of These Risks

AI chatbots are typically trained on data which in actual work could be seen as divulging personal information. Companies utilizing those datasets can find themselves charged expensively with criminal lawsuits or civil penalties.

Sometime in the near future, there will be a growing tide of skepticism bestowed with post-phoniness owing to advancements in technology that impair the use of AI techniques for task accomplishment within the context of consumer expectations of broadening access.

Overview of the PII43K dataset

To create a secure, privacy-aware environment, the PII43K dataset was created and can be an alternative for AI training. Unlike typical available resources, the PII43K allows the training AI models that protect privacy.

Distinctive Features of the PII43K Dataset (Table 1)

PII Masking & Redaction: Replaces PII with placeholders such as [NAME] and [EMAIL], ensuring that no real user information is stored or learned. This also prevents the phenomenon where AI learns to reproduce private information, as broken-up privacy potentials are much more manageable and not requisite unlike raw PII found in multilingual datasets.

Tokenized Data for AI Training: It comes equipped with PII protection as it contains over 43,000 pre-labeled PII entities, which is more than sufficient for AI education in the field of identifying and eliminating sensitive data. The dataset is arranged in templates for exchanging tokens of the data, so the AI should be extracting relevant information.

Enhanced Data Diversity & Contextual Representation: Covers multiple text structures and context variations to improve AI adaptability. AI is trained to detect PII in a formal, informal, or conversational situation.

Better Regulatory Compliance: PII43K does not store private user data, which makes it distinct from usually conceptualized multilingual datasets. It is in agreement with GDPR and CCPA. It eases data issues for organizations in relation to the development of AI-enabled applications.

Improved AI Model Efficiency: Context-bound privacy-filtering entails solid learning capabilities of AI models whereby such models can be applied to the likes of customer service, chatbots, as well as document processing. Less privacy-breaking instances will be found with them than with their models that trained on raw multilingual data.

Optimized for Enterprise and Large-Scale AI Systems: It can be trained to check privacy compliance in auto mode using this database. It supports scalable AI deployments for industries requiring high data privacy, such as finance, healthcare, and legal sectors.

Feature	PII43K (Privacy-Preserving Dataset)	English, French, German, and Italian Datasets
Contains Raw PII	No (Masked & Redacted)	Yes (Exposes Personal Data)
Tracking Risks	No Tracking Data	Yes (Contains IPs, User Agents)
Financial Risks	No Bank Details, Bitcoin, IBANs	Yes (High Exposure)
Privacy-Compliant	Yes (Aligned with GDPR, CCPA)	No (Violates Privacy Laws)
Data Diversity & Contextual Coverage	High (Various Text Scenarios)	Limited (Static Data)
Best Use Case	AI Privacy Research, Secure AI Models	General AI/NLP Training (High Risk)
Risk Level	Low (Privacy-Focused)	High (Data Exposure)

TABLE 1: Comparative Advantage of PII43K Over Multilingual Datasets.

Source: [19]

PII, Personally Identifiable Information; GDPR, General Data Protection Regulation; CCPA, California Consumer Privacy Act; IBAN, International Bank Account Number; AI, Artificial Intelligence; NLP, Natural Language Processing; IPs, Internet Protocols

Discussion

Comparative analysis of privacy risks in multilingual datasets

The data from this study indicates substantial privacy risks in multilingual AI datasets, such as exposure to PII, violation of regulatory compliance, and other safety incidents. Compared to traditional multilingual datasets (English, French, German, and Italian), the use of the privacy-preserving PII43K dataset shows that a very strict level of anonymization must be applied in AI training models.

Key findings

Exposure to PII and Tracking Risks

The presence of raw PII in the multilingual datasets studied, consisting of names, email addresses, financial details, and tracking data of persons' IPs, poses significant privacy threats. With the privileges the datasets offer, they have made unauthorized surveillance and access practically possible, thus creating security loopholes. In contrast, the PII43K dataset has implemented structured anonymization mechanisms ensuring that no direct personal information is stored or processed in raw form.

Risks for Financial Security and Data Protection

In the multilingual datasets, there was severe exposure of very sensitive data, including financial details such as IBANs and cryptocurrency wallet addresses. This raises alarms concerning identity theft and fraud. The PII43K dataset prevents such risks by leaving out such sensitive financial information in the first place, thus fostering security.

Lawful Risks

The increased exposure of financial data in the multilingual datasets, such as IBANs and cryptocurrency wallet addresses, raises the alarm for identity theft or fraudulent activities. The PII43K dataset is said to mitigate these concerns by excluding such sensitive financial information and adding another layer to its safety.

Contextual Coverage and Diversity of the Data

The multilingual datasets are far outside the boundaries of major privacy regulations like the GDPR and CCPA, thereby creating a legal risk for their application in AI training. This means that PII43K meets every requirement under those laws and therefore guarantees privacy-compliant AI development.

Risk Level and Best Use Cases

Analysis shows that multilingual datasets expose sensitive personal data to a high risk, thus making them unsuitable for privacy-sensitive AI applications. On the other hand, being a low-risk dataset, PII43K is

ideal for privacy-related AI research and secure model development.

Limitations

While this study effectively identifies the privacy risks inherent in multilingual AI datasets and investigates in particular the effectiveness of PII43k as an intermediate for the information, it is necessary to note the presence of several gaps. The most considerable of these disparities is the poor dataset assortment. The current research restricts its attention to four languages - English, French, German and Italian - and there emerges an issue of huge perspective blindness as they still suffer from the privacy inadequacy of their AI processes. For example, quite a few languages, particularly those from the minority language areas like indigenous languages, are assumed to have specific data protection weaknesses not tackled in this study.

Also among the limitations is the focus on identifiable information, specifically PII, such as names, emails, finance-related information, and tracking existence. While they may be useful for assessing privacy threats, they are not adequately included in the study as the research only uses direct identifiers and does not consider other elements like writing styles, user behaviors, and geographical location. Another weakness is presented by the fact that research can be carried out on how AI assists in the deduction of private confidential features even after direct personal data has been spaced out even more with time.

The study lastly does not explore how much it would cost, in terms of dollars and computational resources, to implement the approach. A certain aspect of SMD43k puts constraints on the use of raw multilingual data. To lay out structured techniques by an enterprise, AI system might be feasible, but may need significant processing and infrastructure resources as well as contractual, legal, and audit requirements. There is a rich area of future research that deals with the bank for using ‘privacy preserving’ data and the advantages that it will assure when integrated into the AI systems in large settings.

Real-life applications

Systems built on AI drive innovations in data collection and analysis that pose ethical and legal challenges to privacy. The user, not knowing the way their data is processed, becomes a victim of this process. When left unchecked, AI decision-making can wreak havoc on matters that are finely sensitive, such as in health care. Surveillance and predictive analytics bend the will against autonomy and privacy, with algorithmic biases doing the rest. Such discriminatory algorithms being employed for hiring, security, or education reflect the biased value judgments made by training data. A few laws exist, but enforcement has never kept up with AI's unprecedented advances.

AI has somehow randomly set the order of data collection, quietly confronting the choicest issues of illegality and unequal contestation. Predictive analytics can extract confidential information from any data considered common parlance, erasing the line of privacy agreements from tracking and putting in discouragement for user understanding. The reality of continuous data collection hurled disorientation towards tracking yet provides an effective encouragement for better consent mechanism establishment.

The AI technologies and other AI-assisted technologies mentioned above ensure that minimal intrusion upon privacy occurs. Among such technologies and the practices behind them are differential privacy, federated learning, and homomorphic encryption, which protect data without hindering the operational performance of AI. AI directly aids in cyber defense activities through vulnerability detection, threat identification, and fraud prevention. The behavior of users is analyzed to spot anomalies, such as unauthorized logins, in the interest of financial institutions and government networks. Better consent mechanisms and privacy-preserving technologies will stay critical in ensuring that AI is developed responsibly.

Future scope

Hence, future research can build on these findings by taking forward the investigation on the theme of language diversity extending to the analysis of AI datasets. Examination of privacy threats through non-European languages such as Mandarin, Arabic, and Hindi will also provide a more holistic view of the data privacy challenges posed by AI worldwide. Future studies can also take into consideration privacy risks that exist within certain focused datasets, as in the case of medical records, financial transactions, and government databases, where AI-driven decision-making forms an integral role in the protection of user data.

Future-research should continue to develop privacy-preserving AI techniques. Although PII masking, redaction, and structured anonymization have proven to hold some valid safeguards, they should also be combined with and complimented by promising new strategies such as synthetic data generation, adversarial privacy training, and secure AI enclaves that seek to enhance the privacy of AI-consistent models without denting performance. This would involve extensive research on how these methods can be integrated into enterprise AI deployments for scalable, secure AI adoption.

Another promising area for future research is real-world implementation of datasets preserving privacy. Testing the AI application models trained on PII43K in the real-world - in customer service chatbots, legal document processing, and fraud detection systems - can all contribute to the credibility of privacy save in live environments. Long-term monitoring of AI behavior should be researched along with the effect on AI training periods, whether privacy-centric datasets are indeed capable of ignoring long-term PII memorization risks.

Most importantly, changes in regulations will affect the future of AI privacy and security. The evolving laws on AI and compliance measures relevant to privacy should be targeted by future studies in order to maintain the adaptability of the frameworks for privacy-preserving AI. Policy recommendations balancing AI innovation with regulation on privacy are also important factors for building future global governance standards for AI.

Finally, economic feasibility studies can highlight scalability issues for privacy-preserving technologies in AI. Assessing the cost implications of using structured anonymization, encryption techniques, and differential privacy will allow organizations affected to weigh the cost of strengthening AI privacy against the expenses incurred. Research on resource efficiency optimization for privacy-preserving AI will be critical to ensuring secure yet viable AI systems.

Thus, all future research directions can serve to sharpen and improve AI privacy frameworks, so that AI will continue to grow responsibly while, at the same time, protecting user privacy, security of data, and ethical AI development.

Conclusions

This paper describes how very significant privacy risks are actually included in multilingual AI datasets in English, French, German, and Italian. They would typically come from all kinds of genuine data, social media posts, legal documents, customer interactions, and so on and carry huge amounts of personally identifiable information, including names, emails, IP addresses, and financial data, in unredacted forms. The analysis shows there are over 6,49,000 of such privacy tokens, indicating the very high chance of memorization by AI models, identity theft, and regulatory violations.

On the contrary, the PII43K dataset had great privacy-preserving features through systematic masking and redaction achieving ~98% reduction in exposed identifiers and at the same time over 90% of data utility in context. Its performance was also consistent across languages and within privacy laws like GDPR and CCPA. These findings thus make PII43K a crucial asset in training AI systems safely and ethically with scalability into the future concerning responsible development of AI in multilingual settings.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Thenmozhi M, Vaibhav Viswanath, Sahana Naganandh

Critical review of the manuscript for important intellectual content: Thenmozhi M, Vaibhav Viswanath, Sahana Naganandh

Supervision: Thenmozhi M, Vaibhav Viswanath

Acquisition, analysis, or interpretation of data: Vaibhav Viswanath, Sahana Naganandh

Drafting of the manuscript: Vaibhav Viswanath, Sahana Naganandh

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L: The ethics of algorithms: Mapping the debate. Big Data & Society. 2016, 3:10.1177/2053951716679679
2. Jobin A, Ienca M, Vayena E: The global landscape of AI ethics guidelines. Nature Machine Intelligence. 2019, 1:389-399. 10.1038/s42256-019-0088-2
3. Brundage M, Avin S, Wang J, et al.: Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv:2004.07213. 10.48550/arXiv.2004.07213
4. Voigt P, von dem Bussche A: The EU General Data Protection Regulation (GDPR). Springer, Cham; 2017. 10.1007/978-3-319-57959-7
5. Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H: Accountable algorithms. University of Pennsylvania Law Review. 2017, 165:2765268.
6. Brennan-Marquez K, Henderson SE: Artificial intelligence and role-reversible judgment. Faculty Articles and Papers. 2019, 593.
7. Goodfellow IJ, Shlens J, Szegedy C: Explaining and harnessing adversarial examples. arXiv:1412.6572. 10.48550/arXiv.1412.6572
8. Kurakin A, Goodfellow I, Bengio S: Adversarial examples in the physical world. arXiv:1607.0253. 10.48550/arXiv.1607.02533
9. Shokri R, Stronati M, Song C, Shmatikov V: Membership inference attacks against machine learning models. arXiv:1610.05820. 10.48550/arXiv.1610.05820
10. Dwork C, Roth A: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science. 2013, 9:211-407.
11. Doshi-Velez F, Kim B: Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. 10.48550/arXiv.1702.08608
12. Brynjolfsson E, McAfee A: The business of artificial intelligence. Harvard Business Review. 2017, 7:3-11.
13. Kitchin R: Thinking critically about and researching algorithms. Information Communication & Society. 2016, 20:14-29. 10.1080/1369118x.2016.1154087
14. Floridi L: Translating principles into practices of digital ethics: Five risks of being unethical. Philosophy & Technology. 2019, 32:185-193. 10.1007/s13347-019-00354-x
15. Zuboff S: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. Profile Books, London; 2019.
16. Ajunwa I, Crawford K, Schultz J: Limitless worker surveillance. California Law Review. 2016, 105:735-776.
17. Zarsky T: The trouble with algorithmic decisions. Science Technology & Human Values. 2015, 41:118-132. 10.1177/0162243915605575
18. Mutuku M: Legal and ethical implications of data privacy in artificial intelligence: A review of data privacy among learners in Kenyan secondary schools. International Journal of Innovative Science and Research Technology (IJISRT). 2024, 9:537-540. 10.38124/ijisrt/ijisrt24sep208
19. ai4privacy/pii-masking-43k. (2023). Accessed: January 20, 2025: https://huggingface.co/datasets/ai4privacy/pii-masking-43k.